DISPERSIVE AND DISSIPATIVE ERRORS IN THE DPG METHOD WITH SCALED NORMS FOR HELMHOLTZ EQUATION

J. GOPALAKRISHNAN, I. MUGA, AND N. OLIVARES

This paper is dedicated to Leszek Demkowicz on the occasion of his 60$^{\text{th}}$ birthday.

Abstract. We consider the discontinuous Petrov-Galerkin (DPG) method, where the test space is normed by a modified graph norm. The modification scales one of the terms in the graph norm by an arbitrary positive scaling parameter. Studying the application of the method to the Helmholtz equation, we find that better results are obtained, under some circumstances, as the scaling parameter approaches a limiting value. We perform a dispersion analysis on the multiple interacting stencils that form the DPG method. The analysis shows that the discrete wavenumbers of the method are complex, explaining the numerically observed artificial dissipation in the computed wave approximations. Since the DPG method is a nonstandard least-squares Galerkin method, we compare its performance with a standard least-squares method.

1. Introduction

Discontinuous Petrov-Galerkin (DPG) methods were introduced in [8, 10]. The DPG methods minimize a residual norm, so they belong to the class of least-squares Galerkin methods [3, 7, 14], although the functional setting in DPG methods is nonstandard. In this paper, we introduce an arbitrary parameter $\varepsilon > 0$ into the definition of the norm in which the residual is minimized. We study the properties of the resulting family of DPG methods when applied to the Helmholtz equation.

The DPG framework has already been applied to the Helmholtz equation in [12]. An error analysis with optimal error estimates was presented there. There are two major differences in the content of this paper and [12]. The first is the introduction of the above mentioned parameter, ε. When $\varepsilon = 1$, the method here reduces to that in [12]. The use of such scaling parameters was already advocated in [11] based on numerical experience. In this paper, we shall provide a theoretical basis for its use. The second major difference with [12] is that in this contribution we perform a dispersion analysis of the DPG method with the ε scaling. We thus discover several important properties of the method as ε is varied.

Least-squares Galerkin methods are popular methods in scientific computation [3, 17]. They yield Hermitian positive definite systems, notwithstanding the indefiniteness of the underlying problem. Hence they are attractive from the point of view of solver design

2010 *Mathematics Subject Classification.* 65N30,35J05.

Key words and phrases. least-squares, dispersion, dissipation, quasioptimality, resonance, stencil.

This work was partially supported by the NSF under grant DMS-1211635, by the AFOSR under grant FA9550-12-1-0484, and by the FONDECYT project 1110272.

and many works have focused on this subject [18, 19]. However, as we shall shortly see in detail, for wave propagation problems, they yield solutions with heavy artificial dissipation. Since the DPG method is of the least-squares type, it also suffers from this problem. One of the goals of this paper is to show that by means of the ε-scaling, we can rectify this problem to some extent.

To explain this issue further, let us fix the specific boundary value problem we shall consider. Let $A : H(\mathrm{div}, \Omega) \times H^1(\Omega) \to L^2(\Omega)^N \times L^2(\Omega)$ denote the Helmholtz wave operator defined by

$$A(\vec{v}, \eta) = (\hat{\imath}\omega\vec{v} + \vec{\nabla}\eta, \hat{\imath}\omega\eta + \vec{\nabla}\cdot\vec{v}). \tag{1}$$

Here $\hat{\imath}$ denotes the imaginary unit, ω is the wavenumber, and Ω is a bounded open connected domain with Lipschitz boundary. All function spaces in this paper are over the complex field $\mathbb{C}$. The Helmholtz equation takes the form $A(\vec{u}\,, \phi) = \mathsf{f}$, for some $\mathsf{f} \in L^2(\Omega)^N \times L^2(\Omega)$. Although, we consider a general f in this paper, in typical applications, $\mathsf{f} = (\vec{0}, f)$ with $f \in L^2(\Omega)$, in which case, eliminating the vector component $\vec{u}$, we recover the usual second order form of the Helmholtz equation,

$$-\Delta\phi - \omega^2\phi = \hat{\imath}\omega f, \qquad \text{on } \Omega.$$

This must be supplemented with boundary conditions. The DPG method for the case of the impedance boundary conditions $\hat{\imath}\omega\phi + \partial\phi/\partial n = 0$ on $\partial\Omega$ was discussed in [12], but other boundary conditions are equally well admissible. In the present work, we consider the Dirichlet boundary condition

$$\phi = 0, \qquad \text{on } \partial\Omega. \tag{2}$$

To deal with this boundary condition, we will need the space

$$R = H(\mathrm{div}, \Omega) \times H^1_0(\Omega), \tag{3}$$

Thus, our boundary value problem reads as follows:

$$\text{Find } (\vec{u}\,, \phi) \in R \text{ satisfying } A(\vec{u}\,, \phi) = \mathsf{f}. \tag{4}$$

It is well known [16] that except for ω in Σ, an isolated countable set of real values, this problem has a unique solution. We assume henceforth that ω is not in Σ.

Before studying the DPG method for (4), it is instructive to examine the simpler L^2 least-squares Galerkin method. Set $R_h \subset R$ to the Cartesian product of the lowest order Raviart-Thomas and Lagrange spaces, together with the boundary condition in R. The method finds $(\vec{u}\,^{\mathrm{ls}}_h, \phi^{\mathrm{ls}}_h) \in R_h$ such that

$$(\vec{u}\,^{\mathrm{ls}}_h, \phi^{\mathrm{ls}}_h) = \arg\min_{\mathsf{w}\in R_h} \|\mathsf{f} - A\mathsf{w}\|. \tag{5}$$

Throughout, $\|\cdot\|$ denotes the $L^2(\Omega)$ norm, or the natural norm in the Cartesian product of several $L^2(\Omega)$ component spaces. The method (5) belongs to the so-called FOSLS [7] class of methods.

Although (5) appears at first sight to be a reasonable method, computations yield solutions with artificial dissipation. For example, suppose we use (5), appropriately modified

to include nonhomogeneous boundary conditions, to approximate a plane wave propagating at angle $\theta = \pi/8$ in the unit square. A comparison between the real parts of the exact solution (in Figure 1a) and the computed solution (in Figure 1b) shows that the computed solution dissipates at interior mesh points. The same behavior is observed for the lowest order DPG method with $\varepsilon = 1$ in Figure 1c (see §2.4 for the definition of r therein and Section 4 for a full discussion of the lowest order DPG method). The same method with $\varepsilon = 10^{-6}$ however gave a solution (in Figure 1d) that is visually indistinguishable from the exact solution. Note that, for the DPG method with $\varepsilon = 1$, the numerical results presented in [12] show much better performance, because slightly higher order spaces were used there. Instead, in this paper, we have chosen to study the DPG method with the lowest possible order of approximation spaces to reveal the essential difficulties with minimal computational effort.

The situation in Figures 1b and 1c improves when more elements per wavelength are used. This is not surprising in view of the asymptotic error estimates of the methods. To give an example of such an error estimate, consider the case of the impedance boundary conditions considered in [12]. It is proven there that there is a constant $C > 0$, independent of ω and mesh size h, such that the lowest order DPG solution $(\vec{u}_h, \phi_h)$ satisfies

$$\|\vec{u} - \vec{u}_h\| + \|\phi - \phi_h\| \leq C\omega^2 h \tag{6}$$

for a plane wave solution. A critical ingredient in this analysis is the estimate

$$\|w\| \leq C'\|Aw\|, \tag{7}$$

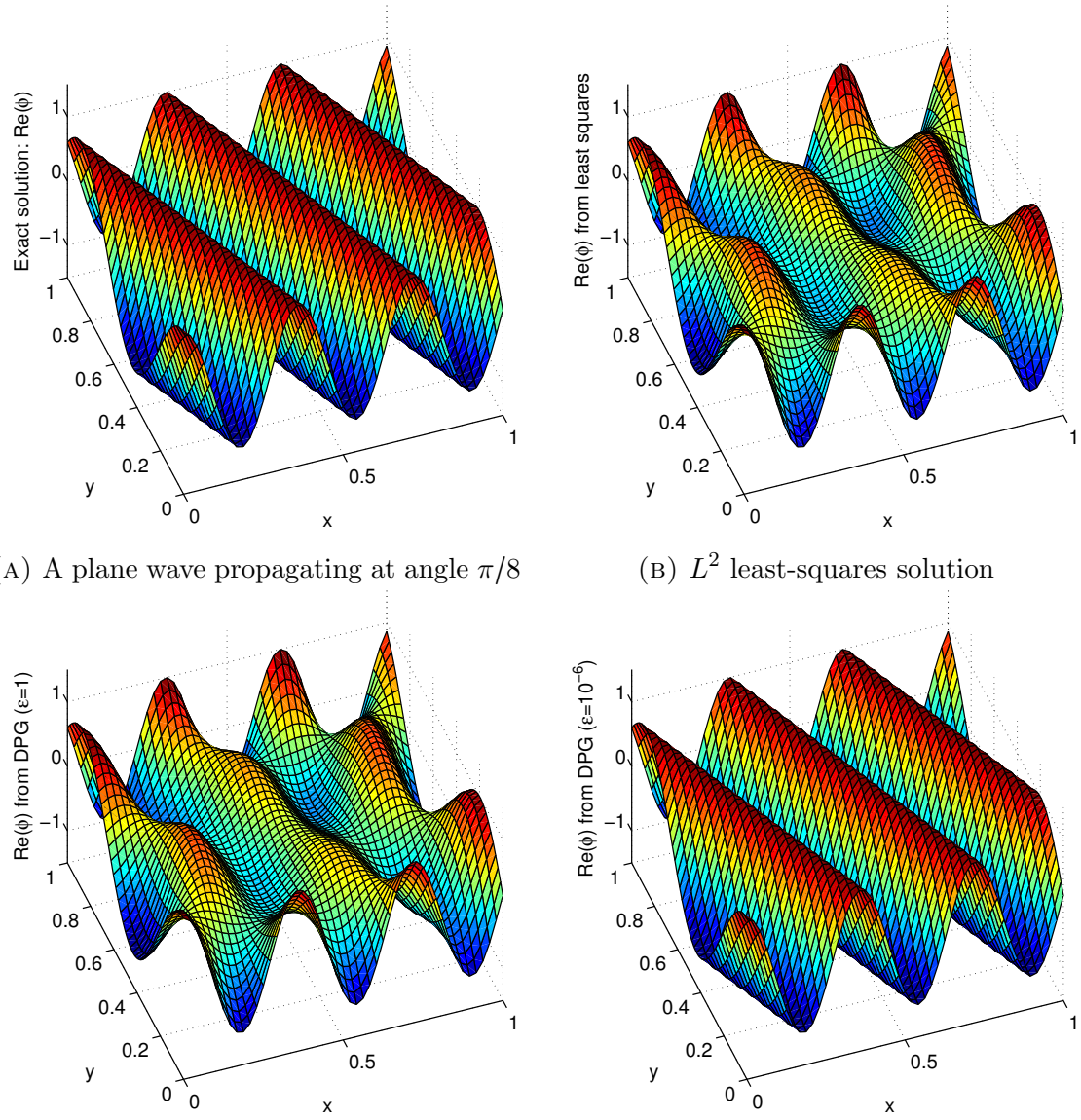
which, as shown in [12, Lemmas 4.2 and 4.3], holds for all w in the analogue of R with impedance boundary conditions. Although the analysis in [12] was for the impedance boundary condition, similar techniques apply to the Dirichlet boundary condition as well, leading to (6). As more elements per wavelength are used, ωh decreases, so (6) guarantees that the situation in Figure 1c will improve.

The analysis for the L^2 least-squares method is easier than the above-mentioned DPG analysis. Indeed, by (5), $\|f - A(\vec{u}_h^{\,\mathrm{ls}}, \phi_h^{\mathrm{ls}})\| \leq \|A(\vec{u} - \vec{w}_h, \phi - \psi_h)\|$ for any $(\vec{w}_h, \psi_h) \in R_h$. Hence, applying (7) to the error $e = (\vec{u} - \vec{u}_h^{\,\mathrm{ls}}, \phi - \phi_h^{\mathrm{ls}})$ and noting that the residual is $Ae = f - A(\vec{u}_h^{\,\mathrm{ls}}, \phi_h^{\mathrm{ls}})$, we obtain $\|e\| \leq C'\|A(\vec{u} - \vec{w}_h, \phi - \psi_h)\|$. By standard approximation estimates, we then conclude that there is a $C > 0$ independent of ω and h such that

$$\|\vec{u} - \vec{u}_h^{\,\mathrm{ls}}\| + \|\phi - \phi_h^{\mathrm{ls}}\| \leq C\omega^2 h. \tag{8}$$

This simple technique of analysis of L^2-based least-squares methods is well-known (see e.g., [17, pp. 70–71]). As with (6), the estimate (8) implies that as the number of elements per wavelength is increased, ωh decreases, and the situation in Figure 1b must improve.

Yet, Figures 1b and 1c show that these methods fail to be competitive with standard methods in accuracy for small number of elements per wavelength. The figures also illustrate one of the difficulties with asymptotic error estimates like (7) and (8). Having little knowledge of the size of C, we cannot predict the performance of the method on coarse meshes. Motivated by this difficulty, one of the theorems we present (Theorem 3.1) will give a better idea of the constant involved as $\varepsilon \to 0$. Also note that the above



(A) A plane wave propagating at angle $\pi/8$

(B) L^2 least-squares solution

(C) Numerical traces from the lowest order DPG method with $\varepsilon = 1$ and $r = 3$

(D) Numerical traces from the lowest order DPG method with $\varepsilon = 10^{-6}$ and $r = 3$

FIGURE 1. Approximations to a plane wave computed using a uniform mesh of square elements of size $h = 1/48$ (about sixteen elements per wavelength). Artificial dissipation is visible in Figures 1b and 1c.

indicated error analyses does not give us a quantitative measure of differences in wave speeds between the computed and exact waves. This motivates the dispersion analysis we present in this paper, which will address the issue of wave speed discrepancies.

We should note that there are alternative methods of the least-squares type that exhibit better performance than the standard L^2-based least squares method. Some are based on adding further terms to the residual to be minimized (e.g., to control the curl of the vector equation [18]). Another avenue explored by others, and closer to the subject

of this paper, is the idea of minimizing the residual in a dual norm [4, 5]. The main difference with our method is that our dual norms are locally computable in contrast to their nonlocal norms. This is achieved by using an ultraweak variational setting. The domain and codomain of the operator in the least-squares minimization associated to the DPG method are nonstandard, as we shall see next.

2. The DPG method for the Helmholtz equation

In this section, we briefly review the method for the Helmholtz equation introduced in [12]. We then show exactly where the parameter ε is introduced to get the variant of the method that we intend to study.

Let Ω_h be a disjoint partitioning of Ω into open elements K such that $\overline{\Omega} = \cup_{K\in\Omega_h}\overline{K}$. The shape of the mesh elements in Ω_h is unimportant for now, except that we require their boundaries ∂K to be Lipschitz so that traces make sense. Let

$$V = H(\mathrm{div},\Omega_h)\times H^1(\Omega_h), \tag{9}$$

where

$$H(\mathrm{div},\Omega_h) = \{\vec{\tau}:\ \vec{\tau}|_K \in H(\mathrm{div},K),\ \forall K\in\Omega_h\},$$
$$H^1(\Omega_h) = \{v:\ v|_K \in H^1(K),\ \forall K\in\Omega_h\}.$$

Let $A_h : V \to L^2(\Omega)^N \times L^2(\Omega)$ be defined in the same way as A in (1), except the derivatives are taken element by element, i.e., on each $K\in\Omega_h$, we have $A_h(\vec{v},\eta)|_K = (\hat{\imath}\omega\vec{v}|_K + \vec{\nabla}\eta|_K, \hat{\imath}\omega\eta|_K + \vec{\nabla}\cdot\vec{v}|_K)$.

2.1. Integration by parts.

The following basic formula that we shall use is obtained simply by integrating by parts each of the derivatives involved:

$$\int_D A(\vec{w},\psi)\cdot\overline{(\vec{v},\eta)} = -\int_D(\vec{w},\psi)\cdot\overline{A(\vec{v},\eta)} + \int_{\partial D}(\vec{w}\cdot\vec{n})\,\overline{\eta} + \int_{\partial D}\psi\,\overline{(\vec{v}\cdot\vec{n})}, \tag{10}$$

for smooth functions $(\vec{w},\psi)$ and $(\vec{v},\eta)$ and domains D with Lipschitz boundary. Above, overlines denote complex conjugations and the integrals use the appropriate Lebesgue measure. Note that we use the notation $\vec{n}$ throughout to generically denote the outward unit normal on various domains – the specific domain will be clear from context – e.g., in (10), it is D. Introducing the following abbreviated notations for tuples $\mathsf{w} = (\vec{w},\psi)$ and $\mathsf{v} = (\vec{v},\eta)$,

$$\langle \mathsf{w},\mathsf{v}\rangle_h = \sum_{K\in\Omega_h}\int_K \vec{w}\cdot\overline{\vec{v}} + \psi\,\overline{\eta},$$
$$\langle\!\langle \mathsf{w},\mathsf{v}\rangle\!\rangle_h = \sum_{K\in\Omega_h}\int_{\partial K}(\vec{w}\cdot\vec{n})\,\overline{\eta} + \int_{\partial K}\psi\,\overline{(\vec{v}\cdot\vec{n})},$$

we can rewrite (10), applied element by element, as

$$\langle A\mathsf{w},\mathsf{v}\rangle_h = -\langle \mathsf{w},A_h\mathsf{v}\rangle_h + \langle\!\langle \mathsf{w},\mathsf{v}\rangle\!\rangle_h. \tag{11}$$

By density, (11) holds for all $\mathsf{w}\in H(\mathrm{div},\Omega)\times H^1(\Omega)$ and all $\mathsf{v}\in V$. Then, $\langle\!\langle\cdot,\cdot\rangle\!\rangle_h$ must be interpreted using the appropriate duality pairing as the last term in (11) contains interelement traces on $\partial\Omega_h = \{\partial K : K\in\Omega_h\}$.

It will be convenient to introduce notation for such traces: Define

$$\mathrm{tr}_h : H(\mathrm{div},\Omega)\times H^1(\Omega) \to \prod_K H^{-1/2}(\partial K)\vec{n}\times H^{1/2}(\partial K)$$

as follows. For any $(\vec{w},\psi)\in H(\mathrm{div},\Omega)\times H^1(\Omega)$, the restriction of $\mathrm{tr}_h(\vec{w},\psi)$ on the boundary of any mesh element ∂K takes the form $((\vec{w}\cdot\vec{n})\vec{n}|_{\partial K},\psi|_{\partial K})\in H^{-1/2}(\partial K)\vec{n}\times H^{1/2}(\partial K)$. Although the meaning of $H^{-1/2}(\partial K)\vec{n}$ is more or less self-evident, to include a proper definition, first let Z denote the space of all functions of the form $\xi\vec{n}$ where ξ is in $H^{1/2}(\partial K)$, normed by $\|\xi\vec{n}\|_Z = \|\xi\|_{H^{1/2}(\partial K)}$. Let Z' denote the dual space of Z. Now, consider the map $M\vec{q} = (\vec{q}\cdot\vec{n})\vec{n}|_{\partial K}$, defined for smooth functions $\vec{q}$ on $\bar{K}$. Since

$$\int_{\partial K} M\vec{q}\cdot\xi\vec{n} = \int_{\partial K}(\vec{q}\cdot\vec{n})\xi$$

(the left and right hand sides extend to duality pairings in Z and $H^{1/2}(\partial K)$, respectively), the standard trace theory implies that M can be extended to a continuous linear operator $M : H(\mathrm{div},K)\to Z'$. The range of M is what we denoted by "$H^{-1/2}(\partial K)\vec{n}$." Throughout this paper, functions in $H^{-1/2}(\partial K)\vec{n}$ appear together with a dot product with $\vec{n}$, so we could equally well consider the standard space $H^{-1/2}(\partial K)$, but the notation simplifies with the former. In particular, with this notation, $\mathrm{tr}_h(\vec{w},\psi)$ is a single-valued function on the element interfaces since $(\vec{w},\psi)$ is globally in $H(\mathrm{div},\Omega)\times H^1(\Omega)$.

2.2. **An ultraweak formulation.** The boundary value problem we wish to approximate is (13). Recall the definition of R in (3). To deal with the Dirichlet boundary condition, we will need the trace space

$$Q = \mathrm{tr}_h(R). \tag{12}$$

To derive the DPG method for

$$A(\vec{u},\phi) = f, \qquad \text{on } \Omega, \tag{13a}$$

$$\phi = 0, \qquad \text{on } \partial\Omega, \tag{13b}$$

we use the integration parts by formula (11) to get

$$-\langle(\vec{u},\phi), A_h(\vec{v},\eta)\rangle_h + \langle\!\langle \mathrm{tr}_h(\vec{u},\phi),(\vec{v},\eta)\rangle\!\rangle_h = \langle f,(\vec{v},\eta)\rangle_h$$

for all $(\vec{v},\eta)\in V$. Now we let the trace $\mathrm{tr}_h(\vec{u},\phi)$ be an independent unknown $(\hat{u},\hat{\phi})$ in Q. Defining the bilinear form $b((\vec{u},\phi,\hat{u},\hat{\phi}),(\vec{v},\eta)) = -\langle(\vec{u},\phi),A_h(\vec{v},\eta)\rangle_h + \langle\!\langle(\hat{u},\hat{\phi}),(\vec{v},\eta)\rangle\!\rangle_h$, we obtain the ultraweak formulation of [12]: Find $u = (\vec{u},\phi,\hat{u},\hat{\phi})$ in

$$U = L^2(\Omega)^N\times L^2(\Omega)\times Q$$

satisfying

$$b(u,v) = \langle f, v\rangle_h, \qquad \forall\, v\in V. \tag{14}$$

The wellposedness of this formulation was proved in [12] for the case of impedance boundary conditions. As is customary, we refer to the solution component $\hat{u}$ as the *numerical flux* and $\hat{\phi}$ as the *numerical trace*.

2.3. **The ε-DPG method.** Let $U_h \subset U$ be a finite dimensional trial space. The *DPG method* finds u_h in U_h satisfying

$$b(u_h, v_h) = \langle f, v_h \rangle_h, \tag{15}$$

for all v_h in the test space V_h, defined by

$$V_h = TU_h, \tag{16}$$

where $T : U \to V$ is defined by

$$\langle Tw, v \rangle_V = b(w, v), \qquad \forall v \in V, \tag{17}$$

and the V-inner product $\langle \cdot, \cdot \rangle_V$ is the inner product generated by the norm

$$\|v\|_V^2 = \|A_h v\|^2 + \varepsilon^2 \|v\|^2. \tag{18}$$

Here $\varepsilon > 0$ is an arbitrary scaling parameter. Note that when $\varepsilon = 1$, (18) defines a *graph norm* on V. The case $\varepsilon = 1$, analyzed in [12], is the standard DPG method. In the next section, we will adapt the analysis of [12] to the case of the variable ε, which we refer to as the "ε-DPG method."

It is easy to reformulate the ε-DPG method as a residual minimization problem. (All DPG methods with test spaces as in (17) minimize a residual as already pointed out in [10].) Letting V' denote the dual space of V, normed with $\|\cdot\|_{V'}$, we define $F \in V'$ by $F(v) = \langle f, v \rangle_h$. Then letting $B : U \to V'$ denote the operator generated by the above-defined $b(\cdot, \cdot)$, i.e., $Bw(v) = b(w, v)$ for all $w \in U$ and $v \in V$, one can immediately see that u_h solves (15) if and only if

$$u_h = \arg \min_{w_h \in U_h} \|Bw_h - F\|_{V'}.$$

This norm highlights the difference between the DPG method and the previously discussed standard L^2-based least-squares method (5).

2.4. **Inexactly computed test spaces.** A basis for the test space V_h, defined in (16), can be obtained by applying T to a basis of U_h. One application of T requires solving (17), which although local (calculable element by element), is still an infinite dimensional problem. Accordingly a practical version of the ε-DPG method uses a finite dimensional subspace $V^r \subset V$ and replaces T by $T^r : U \to V^r$ defined by

$$\langle T^r w, v \rangle_V = b(w, v), \qquad \forall v \in V^r. \tag{19}$$

In computations, we then use, in place of V_h, the inexactly computed test space $V_h^r \equiv T^r U_h$, i.e., the practical DPG method finds u_h^r in U_h satisfying

$$b(u_h^r, v) = \langle f, v \rangle_h, \qquad \forall v \in V_h^r. \tag{20}$$

For the Helmholtz example, we set V^r as follows: Let $\mathcal{Q}_{l,m}$ denote the space of polynomials of degree at most l and m in x_1 and x_2, resp. Let $RT_r \equiv \mathcal{Q}_{r,r-1} \times \mathcal{Q}_{r-1,r}$ denote the Raviart-Thomas subspace of $H(\mathrm{div}, K)$. We set

$$V^r = \{v : v|_K \in RT_r \times \mathcal{Q}_{r,r}\}.$$

Clearly, $V^r \subseteq H(\mathrm{div}, \Omega_h) \times H^1(\Omega_h)$. Later, we shall solve (20) using $r \geq 2$ and report the numerical results. It is easy to see using the Fortin operators developed in [15] that T^r is injective for $r \geq 2$, which implies that (20) yields a positive definite system. However, a complete analysis using [15] tracking ω and r dependencies, remains to be developed, and is not the subject of this paper.

3. Analysis of the ε-DPG method

The purpose of this section is to study how the stability constant of the ε-DPG method (15) depends on ε. The analysis in this section provides the theoretical motivation to introduce the scaling by ε into the DPG setting.

3.1. **Assumption.** The analysis is under the already placed assumption that the boundary value problem (13) is uniquely solvable. We will now need a quantitative form of this assumption. Namely, there is a constant $C(\omega) > 0$, possibly depending on ω, such that the solution of (13) satisfies

$$\|(\vec{u}, \phi)\| \leq C(\omega)\|f\|.$$

One expects $C(\omega)$ to become large as ω approaches any of the resonances in Σ. For any $(\vec{r}, \psi) \in R$, choosing $f = A(\vec{r}, \psi)$ and applying the above inequality, we obtain

$$\|(\vec{r}, \psi)\| \leq C(\omega)\|A(\vec{r}, \psi)\|, \qquad \forall (\vec{r}, \psi) \in R. \tag{21}$$

This is the form in which we will use the assumption.

Note that in the case of the impedance boundary condition, the unique solvability assumption can be easily verified [20] for all ω. Furthermore, when that boundary condition is imposed, for instance, on the boundary of a convex domain, the estimate (21) is proved in [12, Lemmas 4.2 and 4.3] using a result of [20] . The resulting constant $C(\omega)$ is bounded *independently of* ω. However, we cannot expect this independence to hold for the Dirichlet boundary condition (2) we are presently considering.

Finally, let us note that the ensuing analysis applies equally well to the impedance boundary condition: We only need to replace the space R considered here by that in [12] and assume (21) for all functions in the revised R.

3.2. **Quasioptimality.** It is well-known that if there are positive constants C_1 and C_2 such that

$$C_1\|v\|_V \leq \sup_{w \in U} \frac{|b(w, v)|}{\|w\|_U} \leq C_2\|v\|_V, \qquad \forall v \in V, \tag{22}$$

then a quasioptimal error estimate

$$\|u - u_h\|_U \leq \frac{C_2}{C_1} \inf_{w \in U_h} \|u - w\|_U \tag{23}$$

holds. This follows from [12, Theorem 2.1], or from the more general result of [15, Theorem 2.1], after noting that the following uniqueness condition holds: Any $w \in U$ satisfying $b(w, v) = 0$ for all $v \in V$ vanishes. (Since this uniqueness condition can be proved as in [12, Lemma 4.1], we shall not dwell on it here.)

Accordingly, the remainder of this section is devoted to proving (22), tracking the dependence of constants with ε, and using the U-norm we define below. First, let

$$\|(\vec{r},\psi)\|_R = \frac{1}{\varepsilon}\|A(\vec{r},\psi)\|.$$

By virtue of (21), this is clearly a norm under which the space R, defined in (3), is complete. The space Q in (12) is normed by the quotient norm, i.e., for any $\hat{q} \in Q$,

$$\|\hat{q}\|_Q = \inf\{\|r\|_R : \text{ for all } r \in R \text{ such that } \mathrm{tr}_h r = \hat{q}\}.$$

The function in R which achieves the infimum above defines an extension operator $E : Q \to R$ that is a continuous right inverse of tr_h and satisfies

$$\|E\hat{q}\|_R = \|\hat{q}\|_Q. \tag{24}$$

With these notations, we can now define the norm on the trial space by

$$\|(w,\psi,\hat{w},\hat{\psi})\|_U^2 = \|(w,\psi)\|^2 + \|(\hat{w},\hat{\psi})\|_Q^2.$$

The following theorem is proved by extending the ideas in [12] to the ε-DPG method.

Theorem 3.1. *Suppose* (21) *holds and let* $c = C(\omega)\left(C(\omega)\varepsilon/2 + \sqrt{1 + C(\omega)^2\varepsilon^2/4}\right)$. *Then the inf-sup condition in* (22) *holds with* $C_1 = 1/\sqrt{1+c\varepsilon}$ *and the continuity condition in* (22) *holds with* $C_2 = \sqrt{1+c\varepsilon}$. *Hence, the DPG solution admits the error estimate*

$$\|u - u_h\|_U \le (1 + c\varepsilon) \inf_{w \in U_h} \|u - w\|_U.$$

Proof. We first prove the continuity estimate. Let $(w,\hat{q}) \in U$ and let $v \in V$. We use the abbreviated notations $\hat{q} = (\hat{w},\hat{\psi})$, $w = (w,\psi)$, and $v = (\vec{v},\eta)$. By (21) and (24),

$$\|E\hat{q}\| \le C(\omega)\varepsilon\|\hat{q}\|_Q, \qquad \|AE\hat{q}\| = \varepsilon\|\hat{q}\|_Q. \tag{25}$$

The extension E can be used to rewrite $b((w,\hat{q}),v) = -\langle w, A_h v\rangle_h + \langle E\hat{q}, A_h v\rangle_h + \langle AE\hat{q}, v\rangle_h$. Then, applying the Cauchy-Schwarz inequality, and using (25), we have

$$\begin{aligned} |b((w,\hat{q}),v)| &\le \|w\|\|A_h v\| + C(\omega)\varepsilon\|\hat{q}\|_Q\|A_h v\| + \varepsilon\|\hat{q}\|_Q\|v\| \\ &\le \left(\|w\|^2 + \|\hat{q}\|_Q^2\right)^{1/2} t, \end{aligned} \tag{26}$$

where $t^2 = \|A_h v\|^2 + \left(C(\omega)\varepsilon\|A_h v\| + \varepsilon\|v\|\right)^2$. With $a = C(\omega)\varepsilon\|A_h v\|$ and $b = \varepsilon\|v\|$ we apply the inequality $(a+b)^2 \le (1+\alpha^2)a^2 + (1+\alpha^{-2})b^2$ to obtain

$$t^2 \le \left(1 + (1+\alpha^2)C(\omega)^2\varepsilon^2\right)\|A_h v\|^2 + (1+\alpha^{-2})\varepsilon^2\|v\|^2,$$

for any $\alpha > 0$. Setting $\alpha^2 = -1/2 + \sqrt{1/4 + C(\omega)^{-2}\varepsilon^{-2}}$, so that

$$(1+\alpha^2)C(\omega)^2\varepsilon^2 = \alpha^{-2} = c\varepsilon \tag{27}$$

with c as in the statement of the theorem. Hence, $t^2 \le (1+c\varepsilon)\|v\|_V^2$. Returning to (26),

$$|b((w,\hat{q}),v)| \le C_2\|(w,\hat{q})\|_U\|v\|_V.$$

with $C_2 = \sqrt{1+c\varepsilon}$. This verifies the upper inequality of (22).

To prove the lower inequality of (22), let r be the unique function in R satisfying $Ar = v$ for any given $v \in V$. Then, by (21),

$$\|r\| \le C(\omega)\|v\|. \tag{28}$$

Also, since $\|Ar\| = \|v\|$, letting $\hat{r} = \mathrm{tr}_h r$, we have

$$\|\hat{r}\|_Q = \frac{1}{\varepsilon}\|AE\hat{r}\| \le \frac{1}{\varepsilon}\|Ar\| = \frac{1}{\varepsilon}\|v\|. \tag{29}$$

By (11), we have $\langle Ar, v\rangle_h = -\langle r, A_h v\rangle_h + \langle\!\langle \hat{r}, v\rangle\!\rangle_h$, so

$$\|v\|_V^2 = \varepsilon^2\|v\|^2 + \|A_h v\|^2 = \varepsilon^2\, b((z,\hat{r}), v), \tag{30}$$

where $z = r - \varepsilon^{-2}A_h v$, a function that can be bounded using (28), as follows:

$$\begin{aligned}\|z\|^2 &\le (1+\alpha^2)\|r\|^2 + (1+\alpha^{-2})\varepsilon^{-4}\|A_h v\|^2\\ &\le (1+\alpha^2)C(\omega)^2\|v\|^2 + (1+\alpha^{-2})\varepsilon^{-4}\|A_h v\|^2,\end{aligned}$$

for any $\alpha > 0$. Choosing α as in (27) and using (28)–(29),

$$\begin{aligned}\varepsilon^4\|(z,\hat{r})\|_U^2 &= \varepsilon^4\|z\|^2 + \varepsilon^4\|\hat{r}\|_Q^2\\ &\le \left(1 + (1+\alpha^2)C(\omega)^2\varepsilon^2\right)\varepsilon^2\|v\|^2 + (1+\alpha^{-2})\|A_h v\|^2\\ &\le (1+c\,\varepsilon)\left(\varepsilon^2\|v\|^2 + \|A_h v\|^2\right). \qquad (31)\end{aligned}$$

Returning to (30), we now have

$$\|v\|_V^2 = \frac{b((z,\hat{r}),v)}{\|(z,\hat{r})\|_U}\,\varepsilon^2\,\|(z,\hat{r})\|_U \le \left(\sup_{x\in U}\frac{|b(x,v)|}{\|x\|_U}\right)\sqrt{1+c\,\varepsilon}\,\|v\|_V$$

by virtue of (31), verifying the lower inequality of (22) with $C_1 = 1/\sqrt{1+c\,\varepsilon}$. □

Remark 3.2. Although we presented the above result only for the Helmholtz equation, the ideas apply more generally. It seems possible to prove a similar result abstractly, e.g., using the abstract setting in [6], for any DPG application that uses a scaled graph norm analogous to (18) (with the wave operator A_h replaced by suitable others).

3.3. **Discussion.** Theorem 3.1 shows that the use of the ε-scaling in the test norm can ameliorate some stability problems, e.g., those that can arise from large $C(\omega)$.

Observe that the best possible value for the constant C_2/C_1 in (23) is 1. Indeed, if C_2/C_1 equals 1, then the computed solution u_h coincides with the best approximation to u from U_h. Theorem 3.1 shows that the quasioptimality constant of the DPG method approaches the ideal value of 1 as $\varepsilon \to 0$.

However, since the norms depend on ε, we must further examine the components of the error separately, by defining

$$e^2 = \|\vec{u} - \vec{u}_h\|^2 + \|\phi - \phi_h\|^2, \tag{32a}$$

$$\hat{e}^2 = \|AE(\hat{u} - \hat{u}_h, \hat{\phi} - \hat{\phi}_h)\|^2. \tag{32b}$$

The estimate of Theorem 3.1 implies that

$$e^2 + \frac{\hat{e}^2}{\varepsilon^2} \le (1+c\,\varepsilon)^2\left(a^2 + \frac{\hat{a}^2}{\varepsilon^2}\right) \tag{33}$$

where a and $\hat{a}$ are the best approximation errors defined by

$$a^2 = \inf_{(\vec{w},\psi,0,0)\in U_h} \|\vec{u} - \vec{w}\|^2 + \|\phi - \psi\|^2, \tag{34}$$
$$\hat{a}^2 = \inf_{(0,0,\hat{w},\hat{\psi})\in U_h} \|AE(\hat{u} - \hat{w}, \hat{\phi} - \hat{\psi})\|^2.$$

Note that E is independent of ε.

We want to compare the error bounds for the numerical fluxes and traces in the $\varepsilon = 1$ case with the case of $0 < \varepsilon \ll 1$. To distinguish these cases we will denote the error defined in (32b) by $\hat{e}_1$ when $\varepsilon = 1$. Clearly, (33) implies

$$\hat{e}_1^2 \le (1 + c)^2 \left(a^2 + \hat{a}^2\right). \tag{35}$$

For the other case, (33) implies, after multiplying through by ε^2,

$$\hat{e}^2 \le (1 + c\,\varepsilon)^2 \left(\varepsilon^2 a^2 + \hat{a}^2\right).$$

Comparing this with (35), and noting that a and $\hat{a}$ remain the same for different ε, we find that the DPG errors for fluxes and traces admit a *better bound for smaller* ε. Whether the actually observed numerical error improves, will be investigated through the dispersion analysis presented in a later section, as well as in the next subsection.

3.4. **Numerical illustration.** Theorem 3.1 partially explains a numerical observation we now report. We implemented the ε-DPG method by setting the parameter $r = 3$ (see § 2.4) and computed $\mathsf{u}_h^r = (\vec{u}_h^r, \phi_h^r, \hat{u}_h^r, \hat{\phi}_h^r)$. In analogy with (32), define the discretization errors e_r and $\hat{e}_r$ by $e_r^2 = \|\vec{u} - \vec{u}_h^r\|^2 + \|\phi - \phi_h^r\|^2$ and $\hat{e}_r^2 = \|AE(\hat{u} - \hat{u}_h^r, \hat{\phi} - \hat{\phi}_h^r)\|^2$. Although Theorem 3.1 suggests an investigation of

$$\frac{\|\mathsf{u} - \mathsf{u}_h^r\|_U}{\inf_{\mathsf{w}\in U_h} \|\mathsf{u} - \mathsf{w}\|_U} = \left(\frac{e_r^2 + (\hat{e}_r/\varepsilon)^2}{a^2 + (\hat{a}/\varepsilon)^2}\right)^{1/2},$$

due to the difficulty of applying the extension operator E in practice, we have investigated the ratio e_r/a as a function of ω. Recall that a is the $L^2(\Omega)$ best approximation error defined in (34), so e_r/a measures how close the discretization errors are to the best possible.

For a range of wavenumbers ω, we chose the data $\mathsf{f} = (\vec{0}, f)$ so that the exact solution to (13) on the unit square would be $(\vec{u}, \phi) = (-\frac{i}{\omega}\vec{\nabla}\phi, \phi)$, with $\phi = x(1-x)y(1-y)$. Each resulting boundary value problem was then solved using the ε-DPG method with $\varepsilon = 10^{-n}, n = 0, 1, 2, 3, 4$, on a fixed mesh of $h = 1/16$ and the corresponding discretization errors e_r were collected.

The resulting ratios e_r/a are plotted as a function of ω in Figure 2 for a few ε values. First of all, observe that the graph of the ratio begins close to the optimal value of one for all ε values in the figure. Next, observe that the ratio spikes up as ω approaches the exact resonance value $\omega = \pi\sqrt{2} \approx 4.44$, where $C(\omega)$ is infinity. It is interesting to look at the points near (but not at) the resonance. Observe that as ε is decreased, the DPG method exhibits a "regularizing" effect at points near the resonance: E.g., at $\omega = 5$, the values of e_r/a are closer to 1 for smaller ε. It therefore seems advantageous to use smaller ε for problems near resonance.

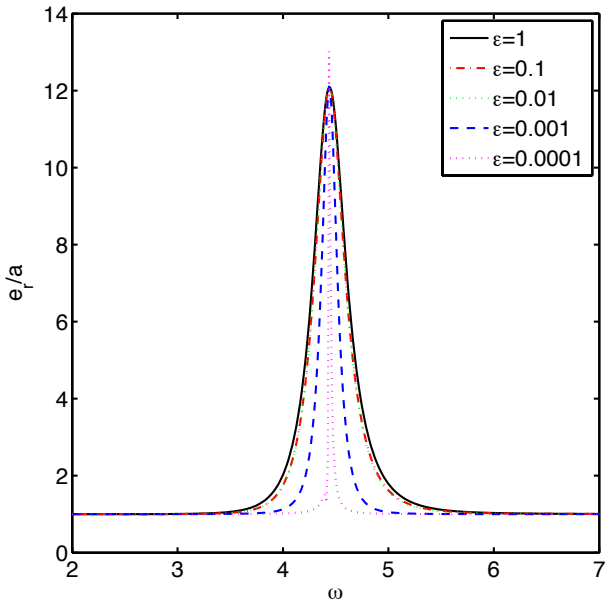


FIGURE 2. The regularizing effect of ε-DPG method as seen from a plot of the ratio e_r/a near a resonance.

The theoretical explanation for this numerical observation would be complete (by virtue of Theorem 3.1), if we had computed using the exact DPG test spaces ($r = \infty$), instead of the inexactly computed spaces ($r = 3$). Certain discrete effects arising due to this inexact computation of test spaces will be presented in a later section.

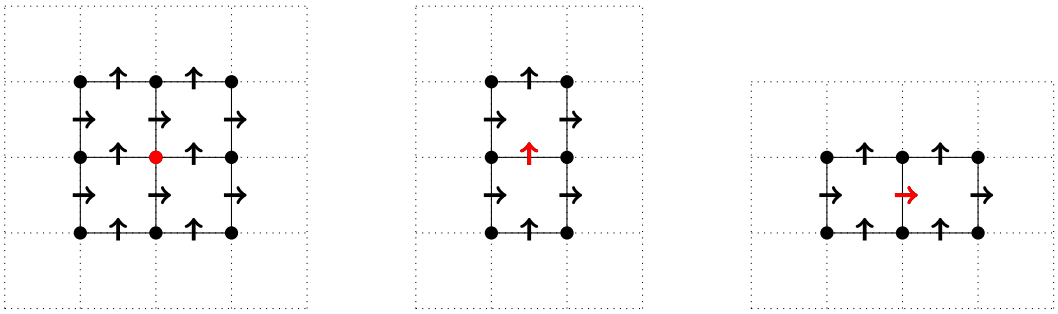
4. Lowest order stencil

We now consider the example of square two-dimensional elements. The lowest order case of the DPG method is obtained using $Q(\partial K) = \{(\hat{w}, \hat{\psi}) : \hat{w}$ is constant on each edge of ∂K, $\hat{\psi}$ is linear on each edge of ∂K, and $\hat{\psi}$ is continuous on $\partial K\}$. Let $S(K) = \{(\vec{w}, \psi) : \vec{w}$ and ψ are constants (vector and scalar, resp.) functions on $K\}$. We consider the DPG method (with ε) using the lowest order global trial space

$$U_h = S_h \times Q_h,$$

where $Q_h = \{\hat{r} \in Q : \hat{r}|_{\partial K} \in Q(\partial K)$ for all mesh elements $K\}$ and $S_h = \{w : w|_K \in S(K)$ for all mesh elements $K\}$.

Let $\hat{\chi}_e$ denote the indicator function of an edge e. If a denotes a vertex of the square element K, let ϕ_a denote the bilinear function that equals one at a and equals zero at the other three vertices of K. Let $\hat{\phi}_a = \phi_a|_{\partial K}$. The collection of eight functions of the form $(0, \hat{\phi}_a)$ and $(\hat{\chi}_e, 0)$, one for each vertex, and one for each edge of K, forms a basis for $Q(\partial K)$. We distinguish between the horizontal and vertical edges, because the unknowns there approximate different components of the velocity $\vec{u}$. Accordingly, we will denote by $\hat{\chi}_e^h$ the indicator function of a horizontal edge and by $\hat{\chi}_e^v$ the indicator function of a vertical edge.



(A) 21-point stencil (B) 13-point stencil (C) 13-point stencil

FIGURE 3. Stencils

The local 11×11 element DPG matrix is defined using a basis for $Q(\partial K)$ and $S(K)$. (While a basis for $Q(\partial K)$ is obtained as mentioned above, a basis for $S(K)$ is trivially obtained by three indicator functions.) If we enumerate the 11 basis functions as e_i, $i = 1, \dots 11$, then the element DPG matrix is defined by

$$B_{ij} = b(e_j, T^r e_i) \tag{36}$$

where T^r is as defined in (19). Since this matrix depends on ω and ε, we will write $B \equiv B(\omega, \varepsilon)$. In our computations, we do not use any specialized basis for V^r to compute the action of T^r, so to overcome round-off problems due to ill-conditioned local matrices, we resorted to high precision arithmetic for these local computations.

To show how B can be computed by mapping, let $\breve{K} = [0,1]^2$. For any square K of side length h, there is a translation vector $\vec{b}_K$ such that the $K - \vec{b}_K = h\breve{K}$. For any (scalar or vector) function v on K, let $\breve{v}$ on $\breve{K}$ be the mapped function obtained by $\breve{v}(\breve{x}) = v(h\breve{x} + \vec{b}_K)$. Let us denote the matrix computed using (36), but using the mapped basis functions $\breve{e}_i$ on $\breve{K}$, by $\breve{B}(\omega, \varepsilon)$. Then by a change of variables, it is easy to see that

$$B(\omega, \varepsilon) = h^2 \breve{B}(\omega h, \varepsilon h). \tag{37}$$

Thus we may compute local DPG matrix by scaling the DPG matrix on the fixed reference element $\breve{K}$ obtained using the *normalized wavenumber* ωh and scaling parameter εh. It is enough to compute the element matrix $\breve{B}$ using high precision arithmetic for the ensuing dispersion analysis.

Next, we eliminate the three interior variables of $S(K)$ and consider the *condensed* 8×8 element stiffness matrix for the variables in $Q(\partial K)$. At this stage it will be useful to classify these eight variables (unknowns) into three categories: (1) Unknowns at vertices a (which are the coefficients multiplying the basis function $\hat{\phi}_a$) denoted by "•", (2) unknowns on horizontal edges (coefficients multiplying $\hat{\chi}^h_e$) denoted by "↑", and (3) unknowns on vertical edges (coefficients multiplying the corresponding $\hat{\chi}^v_e$) denoted by "→". The normal vectors on all horizontal and vertical edges are fixed to be $(0,1)$ and $(1,0)$, respectively, corresponding to the direction of the above-indicated arrows.

Now suppose the mesh is a uniform mesh of congruent square elements. Assembling the above-described condensed 8×8 element matrices on such a mesh, we obtain a global

system where the interior variables are all condensed out. The resulting equations can be represented using the stencils in Figure 3. A row of the matrix system corresponding to an unknown of the type "•", connects to unknowns of the same type at other vertices, as well as unknowns of the other two types, as shown in the 21-point stencil in Figure 3a. Similarly, the unknowns of the type "↑" and "→" connect to other unknowns in the 13-point stencils depicted in Figures 3b and 3c, respectively. These stencils will form the basis of our dispersion analysis next.

5. Dispersion analysis

This section is devoted to a numerical study of the DPG method with ε, by means of a dispersion analysis. The dispersion analysis is motivated by [13]. Details on dispersion analyses applied to standard methods can be found in [1] and the extensive bibliography presented therein.

5.1. **The approach.** To briefly adapt the approach of [13] to fit our context, we consider a general method for the homogeneous Helmholtz equation on an infinite uniform lattice $(h\mathbb{Z})^2$. Suppose the method has S different types of nodes on this lattice, some falling in between the lattice points, each corresponding to a different type of variable, with its own stencil (and hence its own equation). All nodes of the t^{th} type ($t = 1, 2, \ldots, S$) are assumed to be of the form $\vec{\jmath}h$ where $\vec{\jmath}$ lies in an infinite subset of $(\mathbb{Z}/2)^2$. The solution value at a general node $\vec{\jmath}h$ of the t^{th} type is denoted by $\psi_{t,\vec{\jmath}}$. Note that methods with multiple solution components are accommodated using the above mentioned node types.

The t^{th} stencil, centered around $\vec{\jmath}h$, consists of a finite number of nodes, some of which belong to the t^{th} stencil, and the remaining belong to other stencils. Suppose we have finite index sets $J_s \subset (\mathbb{Z}/2)^2$, for each $s = 1, 2, \ldots, S$, such that all the nodes of the t^{th} stencil centered around $\vec{\jmath}h$ can be listed as $N_{\vec{\jmath},t} = \{(\vec{\jmath} + \vec{l}\,)h : \vec{l} \in J_s \text{ and } s = 1, 2, \ldots, S\}$ with the understanding that $(\vec{\jmath} + \vec{l}\,)h$ is a node of s^{th} type whenever $\vec{l} \in J_s$. This allows interaction between variables of multiple types. Every node $(\vec{\jmath}+\vec{l}\,)h$ in $N_{\vec{\jmath},t}$ has a corresponding stencil coefficient (or weight) denoted by $D_{t,s,\vec{l}}$. Due to translational invariance, these weights do not change if we place the stencil at another center node $\vec{\jmath}'h$, hence the numbers $D_{t,s,\vec{l}}$ do not depend on the center index $\vec{\jmath}$.

We obtain the method's equation at a general node $\vec{\jmath}h$ of the t^{th} type by applying the t^{th} stencil centered around $\vec{\jmath}h$ to the solution values $\{\psi_{t,\vec{\jmath}}\}$, namely

$$\sum_{s=1}^{S} \sum_{\vec{l} \in J_s} D_{t,s,\vec{l}}\, \psi_{s,\vec{\jmath}+\vec{l}} = 0. \tag{38}$$

Note that we have set all sources to zero to get a zero right hand side in (38).

Plane waves, $\psi(\vec{x}) \equiv e^{i\vec{k}\cdot\vec{x}}$, are exact solutions of the Helmholtz equation with zero sources (and are often used to represent other solutions). Here the wave vector $\vec{k}$ is of the form $\vec{k} = \omega(\cos(\theta), \sin(\theta))$ for some $0 \le \theta < 2\pi$ representing the direction of propagation. The objective of dispersion analysis is to find similar solutions of the discrete homogeneous

system. Accordingly, we set in (38), the ansatz

$$\psi_{t,\vec{j}} = a_t e^{\hat{\imath}\vec{k}_h\cdot\vec{j}h}, \tag{39}$$

where $\vec{k}_h = \omega_h(\cos(\theta), \sin(\theta))$ and a_t is an arbitrary complex number associated to the t^{th} variable type. We want to find such discrete wavenumbers ω_h satisfying (38).

To this end, we must solve (38) after substituting (39) therein, namely

$$\sum_{s=1}^{S} a_s \sum_{l\in J_s} D_{t,s,\vec{l}}\, e^{\hat{\imath}\vec{k}_h\cdot(\vec{j}+\vec{l})h} = 0, \tag{40}$$

for all $t = 1, 2, \ldots S$. Multiplying by $e^{-\hat{\imath}\vec{k}_h\cdot\vec{j}h}$, we remove any dependence on $\vec{j}$. Defining the $S \times S$ matrix $F \equiv (F_{ts}(\omega_h))$ by

$$F_{ts}(\omega_h) = \sum_{\vec{l}\in J_s} D_{t,s,\vec{l}}\, e^{\hat{\imath}(\omega_h(\cos\theta,\sin\theta)\cdot\vec{l})h},$$

we observe that solving (40) is equivalent to solving

$$\det F(\omega_h) = 0. \tag{41}$$

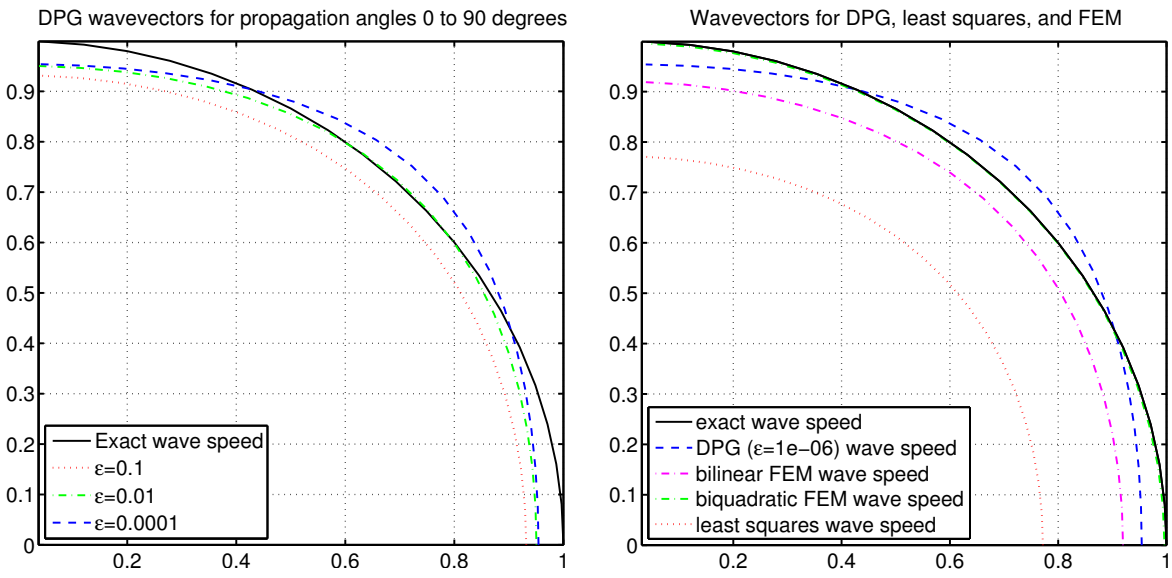
This is the nonlinear equation we solve to obtain the discrete wavenumber ω_h corresponding to any given θ and ω.

5.2. **Application to the DPG method.** Next, we apply the above-described framework to the lowest order DPG stencil discussed in Section 4. Since there are three different types of stencils (see Figure 3), we have $S = 3$. The first type of unknowns, denoted by "$\bullet$", represent the DPG method's approximation to the value of ϕ at the nodes $\vec{j}h$ for all $\vec{j} \in \mathbb{Z}^2$. The stencil of the first type is the one shown in Figure 3a. The unknowns of the second type represent the method's approximation to the vertical components of $\vec{u}$ on the midpoints of horizontal edges, i.e., at all points in $(h\mathbb{Z} + h/2) \times h\mathbb{Z}$. These unknowns were previously denoted by "$\uparrow$" and has the stencil portrayed in Figure 3b. Similarly, the third type of stencil and unknown are as in Figure 3c. To summarize, (39) in the lowest order DPG case, becomes

$$\begin{aligned}
&\psi_{1,\vec{j}} = \hat{\phi}_h(\vec{x}_{\vec{j}}) = a_1 e^{\hat{\imath}\vec{k}_h\cdot\vec{x}_{\vec{j}}} && \forall \vec{x}_{\vec{j}} \in (h\mathbb{Z})^2,\\
&\psi_{2,\vec{j}} = \hat{u}_h(\vec{x}_{\vec{j}}) = a_2 e^{\hat{\imath}\vec{k}_h\cdot\vec{x}_{\vec{j}}} && \forall \vec{x}_{\vec{j}} \in (h\mathbb{Z} + h/2) \times h\mathbb{Z},\\
&\psi_{3,\vec{j}} = \hat{u}_h(\vec{x}_{\vec{j}}) = a_3 e^{\hat{\imath}\vec{k}_h\cdot\vec{x}_{\vec{j}}} && \forall \vec{x}_{\vec{j}} \in h\mathbb{Z} \times (h\mathbb{Z} + h/2).
\end{aligned}$$

The condensed 8×8 DPG matrices, discussed in Section 4, can be used to compute the stencil weights $D_{t,s,\vec{l}}$ in each of the three cases, which in turn lead to the 3×3 nonlinear system (41) for any given propagation angle θ.

We numerically solved the nonlinear system for ω_h, for various choices of θ (propagation angle), r (enrichment degree), ε (scaling factor in the V-norm), and h (mesh size). The first important observation from our computations is that the computed wavenumbers ω_h are complex numbers. They lie close to ω in the complex plane. The small but nonzero imaginary parts of ω_h indicate that the DPG method has dissipation errors, in addition to dispersion errors. The results are described in more detail below.



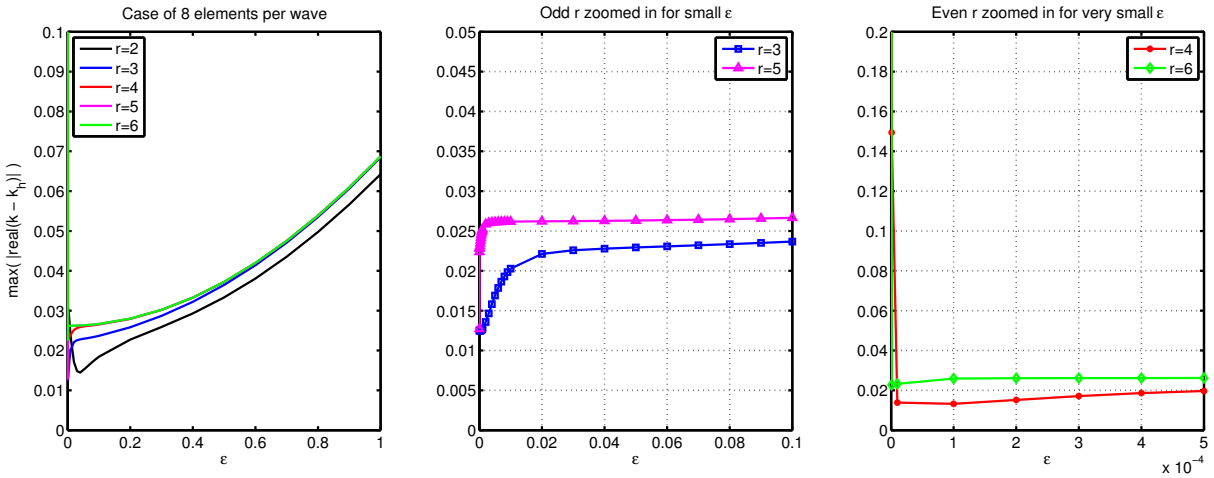
(A) Dependence of $\mathrm{Re}(\vec{k}_h)$ on ε (B) Comparison of methods

FIGURE 4. The curves traced out by the discrete wavevectors $\vec{k}_h$ as θ goes from 0 to $\pi/2$. These plots were obtained using $\omega = 1$ and $h = 2\pi/4$.

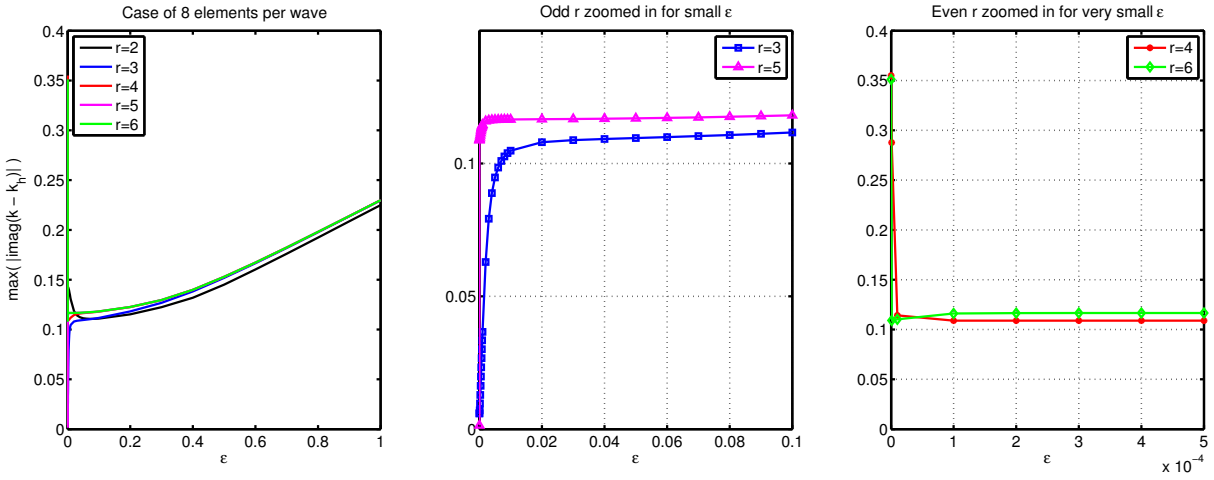
5.3. **Dependence on θ.** To understand how dispersion errors vary with propagation angle θ, we fix the exact wave number ω appearing in the Helmholtz equation to 1 (so the wavelength is 2π) and examine the computed $\mathrm{Re}(\omega_h)$ for each θ.

One way to visualize the results is through a plot of the corresponding discrete wavevectors $\mathrm{Re}(\vec{k}_h)$ vs. $\vec{k}$ for every propagation direction θ. Due to symmetry, we only need to examine this plot in the region $0 \le \theta \le \pi/2$. We present these plots for the case $r = 3$ in Figure 4. We fix $h = 2\pi/4$. (This corresponds to four elements per wavelength if the propagation direction is aligned with a coordinate axis.) In Figure 4a, we plot the curve traced out by the endpoints of the discrete wavevectors $\vec{k}_h$. We see that as ε decreases, the curve gets closer to the (solid) circle traced out by the exact wavevector $\vec{k}$. This indicates better control of dispersive errors with decreasing ε (cf. Theorem 3.1).

In Figure 4b, we compare the $\vec{k}_h$ obtained using the lowest order DPG method with the discrete wavenumbers of the standard lowest order (bilinear) finite element method (FEM). Clearly the wavespeeds obtained from the DPG method are closer to the exact $\omega = 1$ than those obtained by bilinear FEM. However, since the lowest order DPG method has a larger stencil than bilinear FEM, one may argue that a better comparison is with methods having the same stencil size. We therefore compare the DPG method with two other methods which have exactly the same number of points in their stencil: (i) The biquadratic FEM, which after condensation has three stencils of the same size as the lowest order DPG method, and (ii) the conforming first order $L^2(\Omega)$ least-squares method using the lowest order Raviart-Thomas and Lagrange spaces (which has no interior nodes to condense out). While the wavespeeds from the DPG method did not compare favorably



(A) Dispersive errors: Plots of ρ vs. ε



(B) Dissipative errors: Plots of η vs. ε

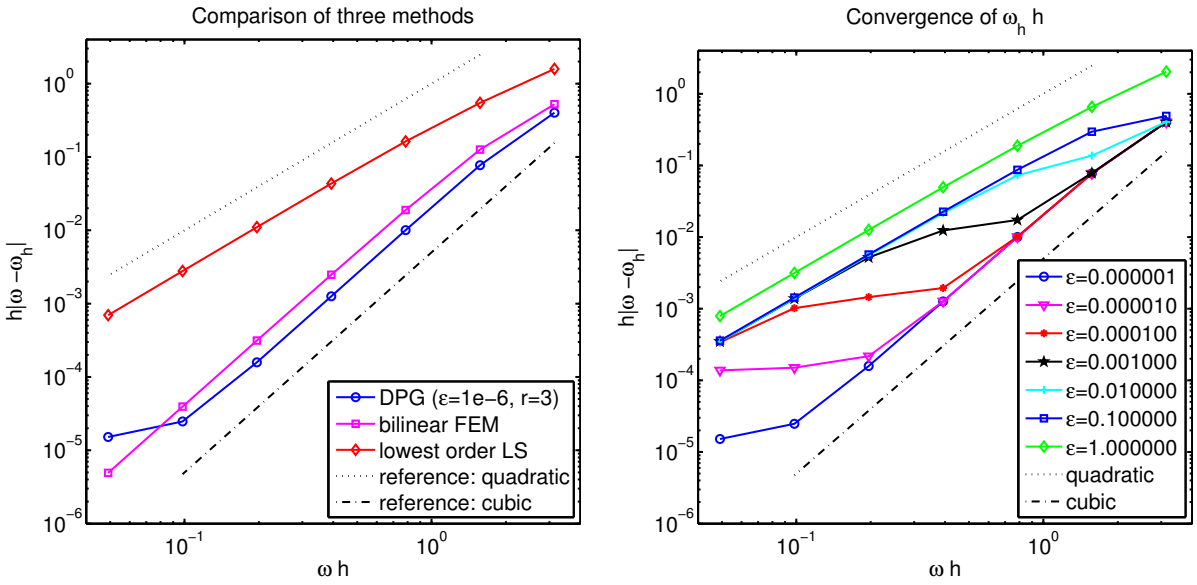
FIGURE 5. The discrepancies between exact and discrete wavenumbers as a function of ε, when $\omega = 1$ and $h = 2\pi/8$.

with the biquadratic FEM of (i), we found that the DPG method performs better than the least-squares method in (ii).

5.4. **Dependence on ε and r.** We have seen in Figure 4 that the discrete wavespeed ω_h is a function of the propagation angle θ. We now examine the maximum discrepancy between real and imaginary parts of ω_h and ω over all angles. Accordingly, define

$$\rho = \max_{\theta} |\mathrm{Re}(\omega_h(\theta)) - \omega| , \qquad \eta = \max_{\theta} |\mathrm{Im}(\omega_h(\theta))| .$$

The former indicates dispersive errors while the latter indicates dissipative errors. Fixing $\omega = 1$ and $h = 2\pi/8$ (so that there are about eight elements per wavelength), we examine these quantities as a function of r and ε in Figure 5. The first of the plots in Figures 5a



(A) Plot of $|\omega_h h - \omega h|$ for three methods (B) Case of DPG with $r = 3$ and various ε

FIGURE 6. Rates of convergence of $|\omega_h h - \omega h|$ to zero for small ωh, in the case of propagation angle $\theta = 0$.

and 5b show that the errors decrease as ε decreases from 1 to about 0.1. In view of Theorem 3.1, we expected this decrease.

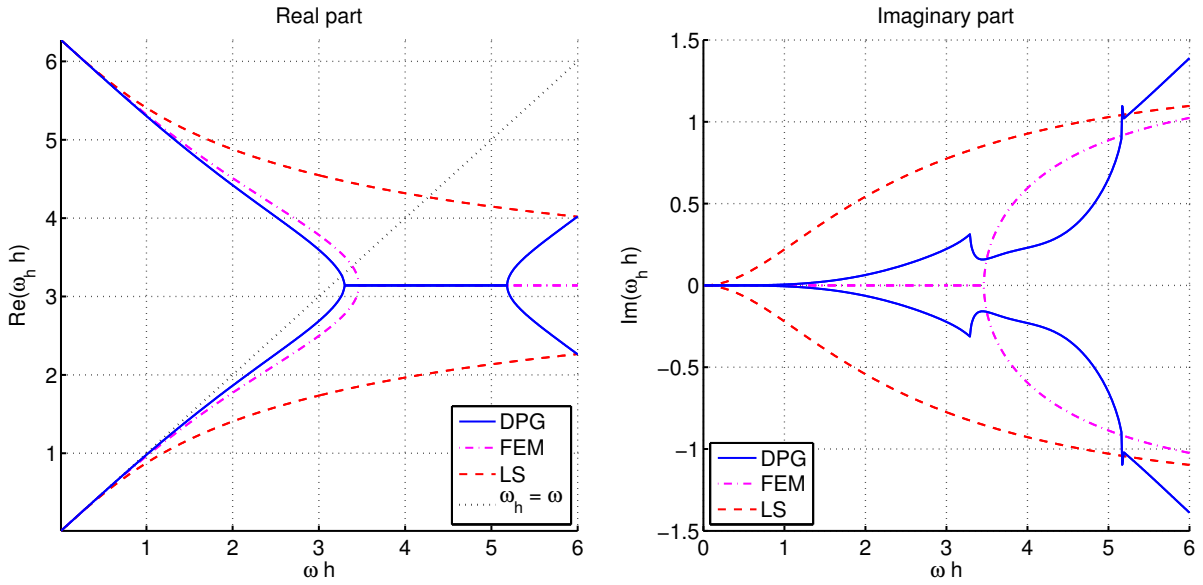
However, the behavior of the method for smaller ε is curious. In the remaining plots of Figure 5 we see that when r is odd, the errors continue to decrease for smaller ε, while for even r, the errors start to increase as $\varepsilon \to 0$. This suggests the presence of discrete effects due to the inexact computation of test functions. We do not yet understand it enough to give a theoretical explanation.

5.5. **Dependence on ω.** Now we examine how ω_h depends on ω. First, let us note that the matrix F in (41) only depends on ωh. (This can be seen, for instance, from (37) and noting how the stencil weights depend on the entries of B.) Hence, we will study how $\omega_h h$ depends on the normalized wavenumber ωh, restricting ourselves to the case of $\theta = 0$.

In Figure 6a, we plot (in logarithmic scale) the absolute value of $\omega_h h - \omega h$ vs. ωh for the standard bilinear FEM, the lowest order L^2 least-squares method (marked LS), and the DPG method with $\varepsilon = 10^{-6}, r = 3$. We observe that while $|\omega_h h - \omega h|$ appears to decrease at $O(\omega h)^2$ for the least squares method, it appears to decrease at the higher rate of $O(\omega h)^3$ for the FEM and DPG cases considered in the same graph. For easy reference, we have also plotted lines indicating slopes corresponding to $O(\omega h)^2$ and $O(\omega h)^3$ decrease, marked "quadratic" and "cubic", resp., in the same graph.

Note that a convergence rate of $|\omega_h h - \omega h| = O(\omega h)^3$ implies that the difference between discrete and exact wave speeds goes to zero at the rate

$$|\omega_h - \omega| = \omega\, O(\omega h)^2.$$



(A) $\mathrm{Re}(\omega_h h)$ as a function of ωh

(B) $\mathrm{Im}(\omega_h h)$ as a function of ωh

FIGURE 7. A comparison of discrete wavenumbers obtained by three lowest order methods in the case of propagation angle $\theta = 0$.

This shows the presence of the so-called [2] pollution errors: For instance, as ω increases, even if we use finer meshes so as to maintain ωh fixed, the error in wave speeds will continue to grow at the rate of $O(\omega)$. Our results show that pollution errors are present in all the three methods considered in Figure 6a. The difference in convergence rates, e.g., whether $|\omega_h - \omega|$ converges to zero at the rate $\omega\, O(\omega h)^2$ or at the rate $\omega\, O(\omega h)$, becomes important, for example, when trying to answer the following question: What h should we use to obtain a fixed error bound for $|\omega_h - \omega|$ for all frequencies ω? While methods with convergence rate $\omega O(\omega h)$ would require $h \approx \omega^{-2}$, methods with convergence rate $\omega O(\omega h)^2$ would only require $h \approx \omega^{-3/2}$.

Next, consider Figure 6b, where we observe interesting differences in convergence rates within the DPG family. While the DPG method for $\varepsilon = 1$ exhibits the same quadratic rate of convergence as the least-squares method, we observe that a transition to higher rates of convergence progressively occur as ε is decreased by each order of magnitude. The $\varepsilon = 10^{-6}$ case shows a rate virtually indistinguishable from the cubic rate in the considered range. The convergence behavior of the DPG method thus seems to vary "in between" those of the least-squares method and the standard FEM as ε is decreased. The values of ωh considered in these plots are $2\pi/2^l$ for $l = 1, 2, \ldots, 7$, which cover the numbers of elements per wavelength in usual practice.

Next, we consider a wider range of ωh following [21], where such a study was done for standard finite elements, separating the real and imaginary parts of $\omega_h h$. Our results for the case of $\theta = 0$ are collected in Figure 7. To discuss these results, let us first recall the behavior of the standard bilinear finite element method (whose discrete wavenumbers are

also plotted in dash-dotted curve in Figure 7). From its well-known dispersion relation (see e.g, [1]), we observe that if $\omega_h h$ solves the dispersion relation, then $2\pi - \omega_h h$ also solves it. Accordingly, the plot in Figure 7a is symmetric about the horizontal line at height π. Furthermore, as already shown in [21], $\omega_h h$ is real-valued in the range $0 < \omega h < \sqrt{12}$. The threshold value $\omega h = \sqrt{12}$ was called the "cut-off" frequency. (Note that in the regime $\omega h > \pi$, we have less than two elements per wavelength. Note also that $\sqrt{12} > \pi$.) As can be seen from Figures 7a and 7b, in the range $\sqrt{12} < \omega h \leq 6$, the bilinear finite elements yield $\omega_h h$ with a constant real part of π and nonzero imaginary parts of increasing magnitude.

We observed a somewhat similar behavior for the DPG method – see the solid curves of Figure 7, which were obtained after calculating F explicitly using the computer algebra package Maple, for the lowest order DPG method, setting $r = 3$ and $\varepsilon = 0$. The major difference between the DPG and FEM results is that $\omega_h h$ from the DPG method was not real-valued even in the regime where FEM wavenumbers were real. It seems difficult to define any useful analogue of the cut-off frequency in this situation. Nonetheless, we observe from Figures 7a and 7b that there is a segment of constant real part of value π, before which the imaginary part of $\omega_h h$ is relatively small. As the number of mesh elements per wavelength increases (i.e., as ωh becomes smaller), the imaginary part of $\omega_h h$ becomes small. We therefore expect the diffusive errors in the DPG method to be small when ωh is small. Finally, we also conclude from Figure 7 that both the dispersive and dissipative errors are better behaved for the DPG method when compared to the L^2 least-squares method.

6. Conclusions

We presented and analyzed the ε-DPG method for the Helmholtz equation. The case $\varepsilon = 1$ was analyzed previously in [12]. The numerical results in [12] showed that in a comparison of the ratio of L^2 norms of the discretization error to the best approximation error is compared, the DPG method had superior properties. The pollution errors reported in [12] for a higher order DPG method were so small that its growth could not be determined conclusively there. In this paper, by performing a dispersion analysis on the DPG method for the lowest possible order, we found that the method has pollution errors that asymptotically grow with ω at the same rate as other comparable methods.

In addition, we found both dispersive and dissipative type of errors in the lowest order DPG method. The dissipative errors manifest in computed solutions as artificial damping of wave amplitudes (e.g., as illustrated in Figure 1).

Our results show that the DPG solutions have higher accuracy than an L^2-based least-squares method with a stencil of identical size. However, the errors in the (lowest order) DPG method did not compare favorably with a standard (higher order) finite element method having a stencil of the same size. Whether this disadvantage can be offset by the other advantages of the DPG methods (such as the regularizing effect of ε, and the fact that it yields Hermitian positive definite linear systems and good gradient approximations) remains to be investigated.

We provided the first theoretical justification for considering the ε-modified DPG method. If the test space were exactly computed, then Theorem 3.1 shows that the errors in numerical fluxes and traces will improve as $\varepsilon \to 0$. However, if the test space is inexactly computed using the enrichment degree r, then the numerical results from the dispersion analysis showed that errors continually decreased as ε was decreased only for odd r. A full theoretical explanation of such discrete effects and the limiting behavior when ε is 0 deserves further study.

References

[1] M. Ainsworth. Discrete dispersion relation for *hp*-version finite element approximation at high wave number. *SIAM J. Numer. Anal.*, 42(2):553–575 (electronic), 2004.

[2] I. M. Babuška and S. A. Sauter. Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers? *SIAM Rev.*, 42(3):451–484 (electronic), 2000.

[3] P. B. Bochev and M. D. Gunzburger. *Least-squares finite element methods*, volume 166 of *Applied Mathematical Sciences*. Springer, New York, 2009.

[4] J. H. Bramble, T. V. Kolev, and J. E. Pasciak. A least-squares approximation method for the time-harmonic Maxwell equations. *J. Numer. Math.*, 13(4):237–263, 2005.

[5] J. H. Bramble, R. D. Lazarov, and J. E. Pasciak. A least-squares approach based on a discrete minus one inner product for first order systems. *Math. Comp.*, 66(219):935–955, 1997.

[6] T. Bui-Thanh, L. Demkowicz, and O. Ghattas. A unified discontinuous Petrov-Galerkin method and its analysis for Friedrichs' systems. *ICES report*, 2011.

[7] Z. Cai, R. Lazarov, T. A. Manteuffel, and S. F. McCormick. First-order system least squares for second-order partial differential equations. I. *SIAM J. Numer. Anal.*, 31(6):1785–1799, 1994.

[8] L. Demkowicz and J. Gopalakrishnan. A class of discontinuous Petrov-Galerkin methods. Part I: The transport equation. *Computer Methods in Applied Mechanics and Engineering*, 199:1558–1572, 2010.

[9] L. Demkowicz and J. Gopalakrishnan. Analysis of the DPG method for the Poisson equation. *SIAM J Numer. Anal.*, 49(5):1788–1809, 2011.

[10] L. Demkowicz and J. Gopalakrishnan, A class of discontinuous Petrov-Galerkin methods. Part II: Optimal test functions, *Numerical Methods for Partial Differential Equations*, 27 (2011), pp. 70–105.

[11] L. Demkowicz, J. Gopalakrishnan, and A. Niemi. A class of discontinuous Petrov-Galerkin methods. Part III: Adaptivity. *Applied Numerical Mathematics*, 62:396–427, 2012.

[12] L. Demkowicz, J. Gopalakrishnan, I. Muga, and J. Zitelli. Wavenumber explicit analysis for a DPG method for the multidimensional Helmholtz equation. *Computer Methods in Applied Mechanics and Engineering*, 213-216:126–138, March 2012.

[13] A. Deraemaeker, I. Babuška, and P. Bouillard. Dispersion and pollution of the FEM solution for the Helmholtz equation in one, two and three dimensions. *International Journal for Numerical Methods in Engineering*, 46:471–499, 1999.

[14] G. J. Fix and M. D. Gunzburger. On numerical methods for acoustic problems. *Comput. Math. Appl.*, 6(2):265–278, 1980.

[15] J. Gopalakrishnan and W. Qiu. An analysis of the practical DPG method. *Math. Comp*, to appear.

[16] F. Ihlenburg. *Finite element analysis of acoustic scattering*, volume 132 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1998.

[17] B.-n. Jiang. *The least-squares finite element method*. Scientific Computation. Springer-Verlag, Berlin, 1998. Theory and applications in computational fluid dynamics and electromagnetics.

[18] B. Lee. First-order system least-squares for elliptic problems with Robin boundary conditions. *SIAM J. Numer. Anal.*, 37(1):70–104 (electronic), 1999.

[19] B. Lee, T. A. Manteuffel, S. F. McCormick, and J. Ruge. First-order system least-squares for the Helmholtz equation. *SIAM J. Sci. Comput.*, 21(5):1927–1949, 2000.
[20] J. M. Melenk. *On Generalized Finite Element Methods*. PhD thesis, University of Maryland, 1995.
[21] L. L. Thompson and P. M. Pinsky. Complex wavenumber Fourier analysis of the p-version finite element method. *J. Computational Mechanics*, 13(4):255–275, 1994.

Portland State University, PO Box 751, Portland, OR 97207-0751, USA.
E-mail address: gjay@pdx.edu

Pontificia Universidad Católica de Valparaíso, Casilla 4059, Valparaíso, Chile.
E-mail address: ignacio.muga@ucv.cl

Portland State University, PO Box 751, Portland, OR 97207-0751, USA.
E-mail address: nmo@pdx.edu